Supporting Information

1. Column description for variant files

1 chr: chromosome number

2 pos(1-based): physical position on the chromosome as to hg38 (1-based coordinate).

For mitochondrial SNV, this position refers to the rCRS (GenBank: NC_012920).

3 ref: reference nucleotide allele (as on the + strand)

4 alt: alternative nucleotide allele (as on the + strand)

5 aaref: reference amino acid

"." if the variant is a splicing site SNP (2bp on each end of an intron)

6 aaalt: alternative amino acid

"." if the variant is a splicing site SNP (2bp on each end of an intron)

7 rs_dbSNP142: rs number from dbSNP 142

8 hg19_chr: chromosome as to hg19, "." means missing

9 hg19_pos(1-based): physical position on the chromosome as to hg19 (1-based coordinate).

For mitochondrial SNV, this position refers to a YRI sequence (GenBank: AF347015)

10 hg18_chr: chromosome as to hg18, "." means missing

11 hg18_pos(1-based): physical position on the chromosome as to hg18 (1-based coordinate)

For mitochondrial SNV, this position refers to a YRI sequence (GenBank: AF347015)

12 genename: gene name; if the nsSNV can be assigned to multiple genes, gene names are

```
      separated by ";"
13    cds_strand: coding sequence (CDS) strand (+ or -)
14    refcodon: reference codon
15    codonpos: position on the codon (1, 2 or 3)
16    codon_degeneracy: degenerate type (0, 2 or 3)
17    Ancestral_allele: the ancestral allele.
      Ancestral alleles of the mitochondrial genome are from RSRS.
      Ancestral alleles of autosomes and X/Y chromosomes are provided by VEP based on
      Ensembl 71. The following comes from its original README file:
      ACTG - high-confidence call, ancestral state supported by the other two sequences
      actg - low-confidence call, ancestral state supported by one sequence only
      N    - failure, the ancestral state is not supported by any other sequence
      -    - the extant species contains an insertion at this position
      .    - no coverage in the alignment
18    AltaiNeandertal: genotype of a deep sequenced Altai Neanderthal
19    Denisova: genotype of a deep sequenced Denisova
20    Ensembl_geneid: Ensembl gene id
21    Ensembl_transcriptid: Ensembl transcript ids (Multiple entries separated by ";")
22    Ensembl_proteinid: Ensembl protein ids
      Multiple entries separated by ";",  corresponding to Ensembl_transcriptids
```

```
23    aapos: amino acid position as to the protein.
            "-1" if the variant is a splicing site SNP (2bp on each end of an intron).
            Multiple entries separated by ";", corresponding to Ensembl_proteinid
24    SIFT_score: SIFT score (SIFTori). Scores range from 0 to 1. The smaller the score the
            more likely the SNP has damaging effect.
            Multiple scores separated by ";", corresponding to Ensembl_proteinid.
25    SIFT_converted_rankscore: SIFTori scores were first converted to SIFTnew=1-SIFTori,
            then ranked among all SIFTnew scores in dbNSFP. The rankscore is the ratio of
            the rank the SIFTnew score over the total number of SIFTnew scores in dbNSFP.
            If there are multiple scores, only the most damaging (largest) rankscore is presented.
            The rankscores range from 0.00963 to 0.91219.
26    SIFT_pred: If SIFTori is smaller than 0.05 (rankscore>0.395) the corresponding nsSNV is
            predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".
            Multiple predictions separated by ";"
27    Uniprot_acc_Polyphen2: Uniprot accession number provided by Polyphen2.
            Multiple entries separated by ";".
28    Uniprot_id_Polyphen2: Uniprot ID numbers corresponding to Uniprot_acc_Polyphen2.
            Multiple entries separated by ";".
29    Uniprot_aapos_Polyphen2: amino acid position as to Uniprot_acc_Polyphen2.
            Multiple entries separated by ";".
```

30 Polyphen2_HDIV_score: Polyphen2 score based on HumDiv, i.e. hdiv_prob.
The score ranges from 0 to 1.
Multiple entries separated by ";", corresponding to Uniprot_acc_Polyphen2.

31 Polyphen2_HDIV_rankscore: Polyphen2 HDIV scores were first ranked among all HDIV scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.02634 to 0.89865.

32 Polyphen2_HDIV_pred: Polyphen2 prediction based on HumDiv, "D" ("probably damaging", HDIV score in [0.957,1] or rankscore in [0.52844,0.89865]), "P" ("possibly damaging", HDIV score in [0.453,0.956] or rankscore in [0.34282,0.52689]) and "B" ("benign", HDIV score in [0,0.452] or rankscore in [0.02634,0.34268]). Score cutoff for binary classification is 0.5 for HDIV score or 0.3528 for rankscore, i.e. the prediction is "neutral" if the HDIV score is smaller than 0.5 (rankscore is smaller than 0.3528), and "deleterious" if the HDIV score is larger than 0.5 (rankscore is larger than 0.3528). Multiple entries are separated by ";".

33 Polyphen2_HVAR_score: Polyphen2 score based on HumVar, i.e. hvar_prob.
The score ranges from 0 to 1.
Multiple entries separated by ";", corresponding to Uniprot_acc_Polyphen2.

34 Polyphen2_HVAR_rankscore: Polyphen2 HVAR scores were first ranked among all HVAR scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of

the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.01257 to 0.97092.

35 Polyphen2_HVAR_pred: Polyphen2 prediction based on HumVar, "D" ("probably damaging", HVAR score in [0.909,1] or rankscore in [0.62797,0.97092]), "P" ("possibly damaging", HVAR in [0.447,0.908] or rankscore in [0.44195,0.62727]) and "B" ("benign", HVAR score in [0,0.446] or rankscore in [0.01257,0.44151]). Score cutoff for binary classification is 0.5 for HVAR score or 0.45833 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5 (rankscore is smaller than 0.45833), and "deleterious" if the HVAR score is larger than 0.5 (rankscore is larger than 0.45833). Multiple entries are separated by ";".

36 LRT_score: The original LRT two-sided p-value (LRTori), ranges from 0 to 1.

37 LRT_converted_rankscore: LRTori scores were first converted as LRTnew=1-LRTori*0.5 if Omega<1, or LRTnew=LRTori*0.5 if Omega>=1. Then LRTnew scores were ranked among all LRTnew scores in dbNSFP. The rankscore is the ratio of the rank over the total number of the scores in dbNSFP. The scores range from 0.00162 to 0.84324.

38 LRT_pred: LRT prediction, D(eleterious), N(eutral) or U(nknown), which is not solely determined by the score.

39 LRT_Omega: estimated nonsynonymous-to-synonymous-rate ratio (Omega, reported by LRT)

40 MutationTaster_score: MutationTaster p-value (MTori), ranges from 0 to 1. Multiple scores are separated by ";". Information on corresponding transcript(s) can

be found by querying http://www.mutationtaster.org/ChrPos.html

41 MutationTaster_converted_rankscore: The MTori scores were first converted: if the prediction is "A" or "D" MTnew=MTori; if the prediction is "N" or "P", MTnew=1-MTori. Then MTnew scores were ranked among all MTnew scores in dbNSFP. If there are multiple scores of a SNV, only the largest MTnew was used in ranking. The rankscore is the ratio of the rank of the score over the total number of MTnew scores in dbNSFP. The scores range from 0.08977 to 0.81031.

42 MutationTaster_pred: MutationTaster prediction, "A" ("disease_causing_automatic"), "D" ("disease_causing"), "N" ("polymorphism") or "P" ("polymorphism_automatic"). The score cutoff between "D" and "N" is 0.5 for MTnew and 0.31709 for the rankscore.

43 MutationTaster_model: MutationTaster prediction models.

44 MutationTaster_AAE: MutationTaster predicted amino acid change.

45 Uniprot_id_MutationAssessor: Uniprot ID number provided by MutationAssessor.

46 Uniprot_variant_MutationAssessor: AA variant as to Uniprot_id_MutationAssessor.

47 MutationAssessor_score: MutationAssessor functional impact combined score (MAori). The score ranges from -5.545 to 5.975 in dbNSFP.

48 MutationAssessor_rankscore: MAori scores were ranked among all MAori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MAori scores in dbNSFP. The scores range from 0 to 1.

49 MutationAssessor_pred: MutationAssessor's functional impact of a variant :

predicted functional, i.e. high ("H") or medium ("M"), or predicted non-functional, i.e. low ("L") or neutral ("N"). The MAori score cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 3.5, 1.9 and 0.8, respectively. The rankscore cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 0.941, 0.61456 and 0.26284, respectively.

50 FATHMM_score: FATHMM default score (weighted for human inherited-disease mutations with Disease Ontology) (FATHMMori). Scores range from -16.13 to 10.64. The smaller the score the more likely the SNP has damaging effect.

Multiple scores separated by ";", corresponding to Ensembl_proteinid.

51 FATHMM_converted_rankscore: FATHMMori scores were first converted to FATHMMnew=1-(FATHMMori+16.13)/26.77, then ranked among all FATHMMnew scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

52 FATHMM_pred: If a FATHMMori score is <=-1.5 (or rankscore >=0.81332) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "T(OLERATED)".

Multiple predictions separated by ";", corresponding to Ensembl_proteinid.

53 PROVEAN_score: PROVEAN score (PROVEANori). Scores range from -14 to 14. The smaller the score the more likely the SNP has damaging effect.

Multiple scores separated by ";", corresponding to Ensembl_proteinid.

54 PROVEAN_converted_rankscore: PROVEANori were first converted to PROVEANnew=1-(PROVEANori+14)/28, then ranked among all PROVEANnew scores in dbNSFP. The rankscore is the ratio of the rank the PROVEANnew score over the total number of PROVEANnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

55 PROVEAN_pred: If PROVEANori <= -2.5 (rankscore>=0.543) the corresponding nsSNV is predicted as "D(amaging)"; otherwise it is predicted as "N(eutral)". Multiple predictions separated by ";", corresponding to Ensembl_proteinid.

56 Transcript_id_VEST3: Transcript id provided by VEST3.

57 Transcript_var_VEST3: amino acid change as to Transcript_id_VEST3.

58 VEST3_score: VEST 3.0 score. Score ranges from 0 to 1. The larger the score the more likely the mutation may cause functional change. Multiple scores separated by ";", corresponding to Transcript_id_VEST3. Please note this score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.

59 VEST3_rankscore: VEST3 scores were ranked among all VEST3 scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of VEST3 scores in dbNSFP. In case there are multiple scores for the same variant, the largest score (most damaging) is presented. The scores range from 0 to 1.

Please note VEST score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.

60 CADD_raw: CADD raw score for functional prediction of a SNP. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5 for details. The larger the score the more likely the SNP has damaging effect. Scores range from -7.535037 to 35.788538 in dbNSFP. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications. For commercial licensing information contact Jennifer McCullar (mccullaj@uw.edu)."

61 CADD_raw_rankscore: CADD raw scores were ranked among all CADD raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of CADD raw scores in dbNSFP. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications. For commercial licensing information contact Jennifer McCullar (mccullaj@uw.edu)."

62 CADD_phred: CADD phred-like score. This is phred-like rank score based on whole genome CADD raw scores. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5

for details. The larger the score the more likely the SNP has damaging effect. Please note the following copyright statement for CADD: "CADD scores (http://cadd.gs.washington.edu/) are Copyright 2013 University of Washington and Hudson-Alpha Institute for Biotechnology (all rights reserved) but are freely available for all academic, non-commercial applications. For commercial licensing information contact Jennifer McCullar (mccullaj@uw.edu)."

63 DANN_score: DANN is a functional prediction score retrained based on the training data of CADD using deep neural network. Scores range from 0 to 1. A larger number indicate a higher probability to be damaging. More information of this score can be found in doi: 10.1093/bioinformatics/btu703. For commercial application of DANN, please contact Daniel Quang (dxquang@uci.edu)

64 DANN_rankscore: DANN scores were ranked among all DANN scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of DANN scores in dbNSFP.

65 fathmm-MKL_coding_score: fathmm-MKL p-values. Scores range from 0 to 1. SNVs with scores >0.5 are predicted to be deleterious, and those <0.5 are predicted to be neutral or benign. Scores close to 0 or 1 are with the highest-confidence. Coding scores are trained using 10 groups of features. More details of the score can be found in doi: 10.1093/bioinformatics/btv009.

66 fathmm-MKL_coding_rankscore: fathmm-MKL coding scores were ranked among all fathmm-MKL coding scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number

of fathmm-MKL coding scores in dbNSFP.

67 fathmm-MKL_coding_pred: If a fathmm-MKL_coding_score is >0.5 (or rankscore >0.28317) the corresponding nsSNV is predicted as "D(AMAGING)"; otherwise it is predicted as "N(EUTRAL)".

68 fathmm-MKL_coding_group: the groups of features (labeled A-J) used to obtained the score. More details can be found in doi: 10.1093/bioinformatics/btv009.

69 MetaSVM_score: Our support vector machine (SVM) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from -2 to 3 in dbNSFP.

70 MetaSVM_rankscore: MetaSVM scores were ranked among all MetaSVM scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaSVM scores in dbNSFP. The scores range from 0 to 1.

71 MetaSVM_pred: Prediction of our SVM based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0. The rankscore cutoff between "D" and "T" is 0.82268.

72 MetaLR_score: Our logistic regression (LR) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging.

Scores range from 0 to 1.

73 MetaLR_rankscore: MetaLR scores were ranked among all MetaLR scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaLR scores in dbNSFP. The scores range from 0 to 1.

74 MetaLR_pred: Prediction of our MetaLR based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.81113.

75 Reliability_index: Number of observed component scores (except the maximum frequency in the 1000 genomes populations) for MetaSVM and MetaLR. Ranges from 1 to 10. As MetaSVM and MetaLR scores are calculated based on imputed data, the less missing component scores, the higher the reliability of the scores and predictions.

76 integrated_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. Integrated (i6) scores are integrated across three cell types (GM12878, H1-hESC and HUVEC). More details can be found in doi:10.1038/ng.3196.

77 integrated_fitCons_rankscore: integrated fitCons scores were ranked among all integrated fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number

of integrated fitCons coding scores in dbNSFP.

78 integrated_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

79 GM12878_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type GM12878. More details can be found in doi:10.1038/ng.3196.

80 GM12878_fitCons_rankscore: GM12878 fitCons scores were ranked among all GM12878 fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GM12878 fitCons coding scores in dbNSFP.

81 GM12878_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

82 H1-hESC_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type H1-hESC. More details can be found in doi:10.1038/ng.3196.

83 H1-hESC_fitCons_rankscore: H1-hESC fitCons scores were ranked among all H1-hESC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of H1-hESC fitCons coding scores in dbNSFP.

84 H1-hESC_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

85 HUVEC_fitCons_score: fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic "fingerprint") that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. GM12878 fitCons scores are based on cell type HUVEC. More details can be found in doi:10.1038/ng.3196.

86 HUVEC_fitCons_rankscore: HUVEC fitCons scores were ranked among all HUVEC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of HUVEC fitCons coding scores in dbNSFP.

87 HUVEC_confidence_value: 0 - highly significant scores (approx. p<.003); 1 - significant scores (approx. p<.05); 2 - informative scores (approx. p<.25); 3 - other scores (approx. p>=.25).

88 GERP++_NR: GERP++ neutral rate

89 GERP++_RS: GERP++ RS score, the larger the score, the more conserved the site. Scores range from -12.3 to 6.17.

90 GERP++_RS_rankscore: GERP++ RS scores were ranked among all GERP++ RS scores in dbNSFP.

The rankscore is the ratio of the rank of the score over the total number of GERP++ RS scores in dbNSFP.

91 phyloP7way_vertebrate: phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 7 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from -5.172 to 1.062 in dbNSFP.

92 phyloP7way_vertebrate_rankscore: phyloP7way_vertebrate scores were ranked among all phyloP7way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP7way_vertebrate scores in dbNSFP.

93 phyloP20way_mammalian: phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from -13.282 to 1.199 in dbNSFP.

94 phyloP20way_mammalian_rankscore: phyloP20way_mammalian scores were ranked among all phyloP20way_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP20way_mammalian scores in dbNSFP.

95 phastCons7way_vertebrate: phastCons conservation score based on the multiple alignments of 7 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.

96 phastCons7way_vertebrate_rankscore: phastCons7way_vertebrate scores were ranked among all phastCons7way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons7way_vertebrate scores in dbNSFP.

97 phastCons20way_mammalian: phastCons conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.

98 phastCons20way_mammalian_rankscore: phastCons20way_mammalian scores were ranked among all phastCons20way_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons20way_mammalian scores in dbNSFP.

99 SiPhy_29way_pi: The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.

100 SiPhy_29way_logOdds: SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site. Scores range from 0 to 37.9718 in dbNSFP.

101 SiPhy_29way_logOdds_rankscore: SiPhy_29way_logOdds scores were ranked among all SiPhy_29way_logOdds scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of SiPhy_29way_logOdds scores in dbNSFP.

102 1000Gp3_AC: Alternative allele counts in the whole 1000 genomes phase 3 (1000Gp3) data.

103 1000Gp3_AF: Alternative allele frequency in the whole 1000Gp3 data.

104 1000Gp3_AFR_AC: Alternative allele counts in the 1000Gp3 African descendent samples.

105 1000Gp3_AFR_AF: Alternative allele frequency in the 1000Gp3 African descendent samples.

106 1000Gp3_EUR_AC: Alternative allele counts in the 1000Gp3 European descendent samples.

107 1000Gp3_EUR_AF: Alternative allele frequency in the 1000Gp3 European descendent samples.

108 1000Gp3_AMR_AC: Alternative allele counts in the 1000Gp3 American descendent samples.

109 1000Gp3_AMR_AF: Alternative allele frequency in the 1000Gp3 American descendent samples.

110 1000Gp3_EAS_AC: Alternative allele counts in the 1000Gp3 East Asian descendent samples.

111 1000Gp3_EAS_AF: Alternative allele frequency in the 1000Gp3 East Asian descendent samples.

112 1000Gp3_SAS_AC: Alternative allele counts in the 1000Gp3 South Asian descendent samples.

113 1000Gp3_SAS_AF: Alternative allele frequency in the 1000Gp3 South Asian descendent samples.

114 TWINSUK_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

115 TWINSUK_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

116 ALSPAC_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

117 ALSPAC_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

118 ESP6500_AA_AC: Alternative allele count in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

119 ESP6500_AA_AF: Alternative allele frequency in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

120 ESP6500_EA_AC: Alternative allele count in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

121 ESP6500_EA_AF: Alternative allele frequency in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

122 ExAC_AC: Allele count in total ExAC samples (~60,706 unrelated individuals)

123 ExAC_AF: Allele frequency in total ExAC samples

124 ExAC_Adj_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC samples

125 ExAC_Adj_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC samples

126 ExAC_AFR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC samples

127 ExAC_AFR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC samples

128 ExAC_AMR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC samples

129 ExAC_AMR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC samples

130 ExAC_EAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC samples

131 ExAC_EAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC samples

132 ExAC_FIN_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC samples

133 ExAC_FIN_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC samples

134 ExAC_NFE_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

135 ExAC_NFE_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

136 ExAC_SAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC samples

137 ExAC_SAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC samples

138 clinvar_rs: rs number from the clinvar data set

139 clinvar_clnsig: clinical significance as to the clinvar data set

2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - drug response,

7 - histocompatibility. A negative score means the the score is for the ref allele

140 clinvar_trait: the trait/disease the clinvar_clnsig referring to

141 Interpro_domain: domain or conserved site on which the variant locates. Domain annotations come from Interpro database. The number in the brackets following a specific domain is the count of times Interpro assigns the variant position to that domain, typically coming from different predicting databases. Multiple entries separated by ";".

2. Column description for gene annotation file

1 Gene_name: Gene symbol from HGNC

2 Ensembl_gene: Ensembl gene id (from HGNC)

3 chr: Chromosome number (from HGNC)

4 Gene_old_names: Old gene symbol (from HGNC)

5 Gene_other_names: Other gene names (from HGNC)

6 Uniprot_acc(HGNC/Uniprot): Uniprot acc number (from HGNC and Uniprot)

7 Uniprot_id(HGNC/Uniprot): Uniprot id (from HGNC and Uniprot)

8 Entrez_gene_id: Entrez gene id (from HGNC)

9 CCDS_id: CCDS id (from HGNC)

10 Refseq_id: Refseq gene id (from HGNC)

11 ucsc_id: UCSC gene id (from HGNC)

12 MIM_id: MIM gene id (from HGNC)

13 Gene_full_name: Gene full name (from HGNC)

14 Pathway(Uniprot): Pathway description from Uniprot

15 Pathway(BioCarta)_short: Short name of the Pathway(s) the gene belongs to (from BioCarta)

16 Pathway(BioCarta)_full: Full name(s) of the Pathway(s) the gene belongs to (from BioCarta)

17 Pathway(ConsensusPathDB): Pathway(s) the gene belongs to (from ConsensusPathDB)

18 Pathway(KEGG)_id: ID(s) of the Pathway(s) the gene belongs to (from KEGG)

19 Pathway(KEGG)_full: Full name(s) of the Pathway(s) the gene belongs to (from KEGG)

20 Function_description: Function description of the gene (from Uniprot)

21 Disease_description: Disease(s) the gene caused or associated with (from Uniprot)

22 MIM_phenotype_id: MIM id(s) of the phenotype the gene caused or associated with (from Uniprot)

23 MIM_disease: MIM disease name(s) with MIM id(s) in "[]" (from Uniprot)

24 Trait_association(GWAS): Trait(s) the gene associated with (from GWAS catalog)

25 GO_biological_process: GO terms for biological process

26 GO_cellular_component: GO terms for cellular component

27 GO_molecular_function: GO terms for molecular function

28 Tissue_specificity(Uniprot): Tissue specificity description from Uniprot

29 Expression(egenetics): Tissues/organs the gene expressed in (egenetics data from BioMart)

30 Expression(GNF/Atlas): Tissues/organs the gene expressed in (GNF/Atlas data from BioMart)

31 Interactions(IntAct): The number of other genes this gene interacting with (from IntAct).

Full information (gene name followed by Pubmed id in "[]") can be found in the ".complete" table

32 Interactions(BioGRID): The number of other genes this gene interacting with (from BioGRID)
Full information (gene name followed by Pubmed id in "[]") can be found in the ".complete" table

33 Interactions(ConsensusPathDB): The number of other genes this gene interacting with (from ConsensusPathDB). Full information (gene name followed by Pubmed id in "[]") can be found in the ".complete" table

34 P(HI): Estimated probability of haploinsufficiency of the gene
(from doi:10.1371/journal.pgen.1001154)

35 P(rec): Estimated probability that gene is a recessive disease gene
(from DOI:10.1126/science.1215040)

36 Known_rec_info: Known recessive status of the gene (from DOI:10.1126/science.1215040)
"lof-tolerant = seen in homozygous state in at least one 1000G individual"
"recessive = known OMIM recessive disease"
(original annotations from DOI:10.1126/science.1215040)

37 RVIS: Residual Variation Intolerance Score, a measure of intolerance of mutational burden,
the higher the score the more tolerant to mutational burden the gene is.
from doi:10.1371/journal.pgen.1003709

38 RVIS_percentile: The percentile rank of the gene based on RVIS, the higher the percentile
the more tolerant to mutational burden the gene is.

39 Essential_gene: Essential ("E") or Non-essential phenotype-changing ("N") based on Mouse Genome Informatics database. from doi:10.1371/journal.pgen.1003484

40 MGI_mouse_gene: Homolog mouse gene name from MGI

41 MGI_mouse_phenotype: Phenotype description for the homolog mouse gene from MGI

42 ZFIN_zebrafish_gene: Homolog zebrafish gene name from ZFIN

43 ZFIN_zebrafish_structure: Affected structure of the homolog zebrafish gene from ZFIN

44 ZFIN_zebrafish_phenotype_quality: Phenotype description for the homolog zebrafish gene from ZFIN

45 ZFIN_zebrafish_phenotype_tag: Phenotype tag for the homolog zebrafish gene from ZFIN

3. Column description for dbscSNV files

1 chr: chromosome number

2 pos: physical position on the chromosome as to hg19 (1-based coordinate)

3 ref: reference nucleotide allele (as on the + strand)

4 alt: alternative nucleotide allele (as on the + strand)

5 hg38_chr: chromosome number as to hg38

6 hg38_pos: physical position on the chromosome as to hg38 (1-based coordinate)

7 RefSeq?: whether the SNV is a scSNV according to RefSeq

8 Ensembl?: whether the SNV is a scSNV according to Ensembl

9 RefSeq_region: functional region the SNV located according to RefSeq

10 RefSeq_gene: gene name according to RefSeq

11 RefSeq_functional_consequence: functional consequence of the SNV according to RefSeq

12 RefSeq_id_c.change_p.change: SNV in format of c.change and p.change according to RefSeq

13 Ensembl_region: functional region the SNV located according to Ensembl

14 Ensembl_gene: gene id according to Ensembl

15 Ensembl_functional_consequence: functional consequence of the SNV according to Ensembl

16 Ensembl_id_c.change_p.change: SNV in format of c.change and p.change according to Ensembl

17 ada_score: ensemble prediction score based on ada-boost. Ranges 0 to 1. The larger the score the higher probability the scSNV will affect splicing. The suggested cutoff for a binary prediction (affecting splicing vs. not affecting splicing) is 0.6.

18 rf_score: ensemble prediction score based on random forests. Ranges 0 to 1. The larger the score the higher probability the scSNV will affect splicing. The suggested cutoff for a binary prediction (affecting splicing vs. not affecting splicing) is 0.6.